## **Scraping Data from a Website**

```
In [4]: from bs4 import BeautifulSoup
         import requests
In [30]:
         # get the url of the webpage
         url = "https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_5
         #import the webpage to the python shell into a html doctype and print it
         page = requests.get(url)
         soup = BeautifulSoup(page.text, 'html')
         print(soup)
         <!DOCTYPE html>
         <html class="client-nojs vector-feature-language-in-header-enabled vector-f</pre>
         eature-language-in-main-page-header-disabled vector-feature-sticky-header-d
         isabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned
         -enabled vector-feature-main-menu-pinned-disabled vector-feature-limited-wi
         dth-enabled vector-feature-limited-width-content-enabled vector-feature-zeb
         ra-design-disabled dir="ltr" lang="en">
         <head>
         <meta charset="utf-8"/>
         <title>List of largest companies in the United States by revenue - Wikipedi
         <script>document.documentElement.className="client-js vector-feature-langua
         ge-in-header-enabled vector-feature-language-in-main-page-header-disabled v
         ector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disab
         led vector-feature-toc-pinned-enabled vector-feature-main-menu-pinned-disab
         led vector-feature-limited-width-enabled vector-feature-limited-width-conte
         nt-enabled vector-feature-zebra-design-disabled";(function(){var cookie=doc
         ument.cookie.match(/(?:^|; )enwikimwclientprefs=([^;]+)/);if(cookie){var fe
         atureName=cookie[1];document.documentElement.className=document.documentEle
                                                 - L.T. J.L. C. E. K. . . L.J. J.Y. Y.Y.
```

```
In [14]: # extract the table concerned from the three tables contained in the entire well
        table = soup.find_all('table')[1]
        print(table)
        <caption>
        </caption>
        Rank
        Name
        Industry
        Revenue <br/> (USD millions)
        Revenue growth
        Employees
        Headquarters
        In [32]:
       # get the table headers
        head = table.find all('th')
        print(head)
        [Rank
        , Name
        , Industry
        , Revenue <br/>(USD millions)
        , Revenue growth
        , Employees
        , Headquarters
        ]
In [31]: # create list coprehension containing the table headers cleaned up
        world titles = [headers.text.strip() for headers in head]
        print(world titles)
        ['Rank', 'Name', 'Industry', 'Revenue (USD millions)', 'Revenue growth', 'Emp
        loyees', 'Headquarters']
In [21]: import pandas as pd
        df = pd.DataFrame(columns= world_titles)
       df
Out[21]:
          Rank Name Industry Revenue (USD millions) Revenue growth Employees Headquarters
```

localhost:8888/notebooks/Documents/first\_Web\_scrapping.ipynb

```
In [29]: #find all the table rows
    all_rows = table.find_all('tr')

for data in all_rows[1:]:

    #extract each table data
    row_data = data.find_all('td')
    list_row_data = [items.text.strip() for items in row_data]

# appending each row to the newly created dataframe
    index = df.shape[0]
    df.loc[index] = list_row_data

display(df)
```

Rank		Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters
0	1	Walmart	Retail	611,289	6.7%	2,100,000	Bentonville, Arkansas
1	2	Amazon	Retail and Cloud Computing	513,983	9.4%	1,540,000	Seattle, Washington
2	3	Exxon Mobil	Petroleum industry	413,680	44.8%	62,000	Spring, Texas
3	4	Apple	Electronics industry	394,328	7.8%	164,000	Cupertino, California
4	5	UnitedHealth Group	Healthcare	324,162	12.7%	400,000	Minnetonka, Minnesota
195	96	Best Buy	Retail	46,298	10.6%	71,100	Richfield, Minnesota
196	97	Bristol-Myers Squibb	Pharmaceutical industry	46,159	0.5%	34,300	New York City, New York
197	98	United Airlines	Airline	44,955	82.5%	92,795	Chicago, Illinois
198	99	Thermo Fisher Scientific	Laboratory instruments	44,915	14.5%	130,000	Waltham, Massachusetts
199	100	Qualcomm	Technology	44,200	31.7%	51,000	San Diego, California

200 rows × 7 columns

```
In [34]: # save this table as a csv file
df.to_csv(r'C:\Users\Meli Imelda\Documents\first_web_scraping_output\companies
```

In	[	]:	
In	[	]:	
In	[	]:	
In	[	]:	
In	[	]:	
In	[	]:	
In	[	]:	